(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 291 848 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
12.03.2003 Bulletin 2003/11

(51) Int Cl.[7]: **G10L 13/08**, G10L 15/14

(21) Application number: **02255451.3**

(22) Date of filing: **05.08.2002**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **31.08.2001 US 942609**

(71) Applicant: **Nokia Corporation**
02150 Espoo (FI)

(72) Inventors:
- **Riis, Soeren**
  2750 Ballerup (DK)
- **Jensen, Kare Jean**
  2830 Virum (DK)
- **Pedersen, Morten With**
  2100 Koebenhavn (DK)

(74) Representative: **Geary, Stuart Lloyd et al**
Venner, Shipley & Co.,
20 Little Britain
London EC1A 7DH (GB)

(54) Multilingual pronunciations for speech recognition

(57) There is provided a novel approach for generating multilingual text-to-phoneme mappings for use in multilingual speech recognition systems. The multilingual mappings are based on the weighted outputs from a neural network text-to-phoneme model, trained on data mixed from several languages. The multilingual mappings used together with a branched grammar decoding scheme is able to capture both inter- and intra-language pronunciation variations which is ideal for multilingual speaker independent speech recognition systems. A significant improvement in overall system performance is obtained for a multilingual speaker independent name dialling task when applying multilingual instead of language dependent text-to-phoneme mapping.

EP 1 291 848 A2

Description

[0001] The invention relates to a method of speech recognition comprising receiving an acoustic input and a speech recognition apparatus comprising transducer means for receiving an acoustic input and processing means.

[0002] Speaker independent command word recognition and name dialling on portable devices such as mobile phones and personal digital assistants has attracted significant interest recently. A speech recognition system provides an alternative to keypad input for limited size portable products. The speaker independence makes the system particularly attractive from a user point of view compared to speaker dependent systems. For large vocabularies, user training of a speaker dependent recogniser is likely to become too tedious to be useful.

[0003] How to build acoustic models that integrate multiple languages in automatic speech recognition applications is described by F. Palou, P. Bravetti, O. Emem, V. Fischer, and E. Janke, in the publication "Towards a Common Phone Alphabet for Multilingual Speech Recognition", In *Proceedings of ICSLP*, pages 1—1,2000.

[0004] An architecture for embedded multilingual speech recognition systems is proposed by O. Viiki, I. Kiss, and J. Tian, in the publication "Speaker- and Language-independent Speech Recognition in Mobile Communication Systems". in *Proceedings of ICASSP*, 2001.

[0005] Use of neural networks for TTP giving estimates of the posterior probabilities of the different phonemes for each letter input is taught by K. Jensen, and S. Riis, in the publication "Self-Organizing Letter Code-Book for Text-To-Phoneme Neural Network Model", published in *Proceedings of ICSLP*, 2000.

[0006] A phoneme-based speaker independent system is ready to use "out-of-the-box" and does not require any training session of the speaker. Furthermore, if the phoneme based recogniser is combined with a text-to-phoneme (TTP) mapping module for generating phoneme pronunciations online from written text, the user may define specific vocabularies as required in e.g. a name dialling application. Naturally, speaker and vocabulary independence comes at a cost, namely increased complexity for real-time decoding, increased requirements for model storage, and usually also a slight drop in recognition performance compared to speaker dependent systems. Furthermore, speaker independent systems typically contain a number of language dependent modules, e.g. language dependent acoustic phoneme models, TTP modules etc. For portable devices, the support of several languages may be prohibited by the limited memory available in such devices as separate modules need to be stored for each language.

[0007] Recently, systems based on multilingual acoustic phoneme models have emerged — see the letters written by Palou et al., and Viikiet al., mentioned above. These systems are designed to handle several different languages simultaneously and are based on the observation that many phonemes are shared among different languages. The basic idea in multilingual acoustic modelling is to estimate the parameters of a particular phoneme model using speech data from all supported languages that include this phoneme. Multilingual speech recognition is very attractive as it makes a particular speech recognition application usable by a much wider audience. In addition the logistic needs is reduced when making world wide products. Furthermore, sharing of phoneme models across languages can significantly reduce memory requirements compared to using separate models for each language. Multilingual recognisers are thus very attractive for portable platforms with limited resources.

[0008] Even though multilingual acoustic modelling has proven efficient, user definable vocabularies typically still require language dependent TTP modules for each supported language. Prior to running the language dependent TTP module it is furthermore necessary to first identify the language ID of each vocabulary entry.

Language Dependent text-to-phoneme (TTP) Mapping

[0009] For applications like speaker independent name dialling on mobile phones the vocabulary entries are typically names in the phonebook database 21 that may be changed at any time. Thus, for a multilingual speaker independent name dialler 22 to work with language dependent TTP, a language identification module (LID) 30 is needed. An example of a multilingual speech recognition system according to prior art using a LID module 30 is shown in Figure 3. In Figure 3, it is shown how the LID module 30 selects a language dependent TTP module 31.1-31.n that is used for generating the pronunciation by means of a pronunciation lexicon module 32 for a multilingual recogniser 33 based on multilingual acoustic phoneme models.

[0010] The LID module 30 may be a statistical model predicting the language ID of each entry in the vocabulary, a deterministic module that sets the language ID of each entry based on application specific knowledge, a module that simply requires the user to set the language ID manually or a combination of these. In the most general case, a priori knowledge about the language ID is not available and manual language identification by the user may not be desirable. In that case, language identification must be based on a statistical LID module that predicts the language ID from the written text.

[0011] Depending on the application, the TTP module 31.1-31.n may be a statistical model, a rule based model, based on a lookup table that contains all possible words, or any combination of these. The latter approach will typically not be possible for name dialling applications on portable devices with limited memory resources, due to the large

number of possible names.

[0012] In most applications based on user defined vocabularies, a statistical LID module 30 has very limited text data for deciding the language ID of an entry. For e.g. a short name like "Peter", only five letters are available for language identification. Furthermore, many names are not unique for a single language but rather used in a large number of languages with different pronunciations. In addition to this, a speaker may pronounce a foreign/non-native name with a significant accent, i.e., the pronunciation of the name is actually a mixture of the pronunciation corresponding to the language from which the name originates and the native language of the speaker.

[0013] This implies that the combination of language dependent TTP modules 31.1-31.n, a statistical LID module 30 and multilingual acoustic phoneme models is likely to give a poor overall performance. Furthermore, if several languages are to be supported in a portable device, the size of the LID and TTP modules may have to be severely limited in order to fit into the low memory resources of the device. For "irregular" languages, like English, high accuracy TTP modules may take up as much as 40—300 kb of memory, whereas TTP modules for rule based "regular" languages like Japanese and Finnish typically require less than 1 kb.

[0014] A method according to the present invention is characterised by generating sequences of multilingual phoneme symbols based on a semantic code or code sequence, generating reference speech signals in dependence on said sequences of multilingual phoneme symbols and comparing said reference speech signals with the acoustic input in order to find a match.

[0015] A speech recognition apparatus according to the present invention is characterised in that the processing means is configured for generating sequences of multilingual phoneme symbols based on a semantic code or code sequence, generating reference speech signals in dependence on said sequences of multilingual phoneme symbols and comparing said reference speech signals with the acoustic input in order to find a match.

[0016] The semantic code or code sequence may represent characters of a known writing system, such as Roman, Cyrillic and Korean alphabets, Japanese syllabaries and Chinese ideographs.

[0017] Where, the semantic code or code sequence represents one or more characters of an alphabet or a syllabary, it is preferable that the code or code sequence be processed character by character and a neural network provide an estimate of the posterior probabilities of the different phonemes or phoneme sequences for each character.

[0018] Preferably, a user input is received, a semantic code or code sequence is generated in dependence on said user input and stored in a store, and the stored code or code sequence is retrieved for generation of said reference speech signals.

[0019] Preferably, the neural network comprises a standard fully connected feed-forward multi-layer perceptron neural network.

[0020] An apparatus according to the present invention may be advantageously incorporated into a communication terminal, for example a mobile phone.

[0021] In the case of a communication terminal, the store comprises data forming an electronic phonebook including phone numbers and associated name labels.

[0022] An embodiment of the present invention will now be described, by way of example only, with reference to the accompanying drawings, in which:-

Figure 1 schematically illustrates a preferred embodiment of a hand portable phone according to the invention.
Figure 2 schematically shows the essential parts of a telephone for communication with e.g. a cellular network.
Figure 3 shows as a block diagram a multilingual speech recognition system employing a LID module and language specific TTP modules according to prior art.
Figure 4 shows as a block diagram a multilingual speech recognition system employing a Multilingual TTP module according to a preferred embodiment of the invention.
Figure 5 shows a branching diagram according to the preferred embodiment of the invention for pronunciation of name "Peter" in German (p-e:-t-6), English (o-i:-t-@) and Spanish (p-i-t-e-r) arranged as a branched grammar. The values on the arcs between phonemes indicate probabilities of the phonemes as provided by e.g. the TTP module. SAMPA notation is used for phonemes.

[0023] Figure 1 shows a preferred embodiment of a phone according to the invention, and it will be seen that the phone, which is generally designated by 1, comprises a user interface having a keypad 2, a display 3, an on/off button 4, a speaker 5 (only openings are shown), and a microphone 6 (only openings are shown). The phone 1 according to the preferred embodiment is adapted for communication preferable via a cellular network e.g. GSM.

[0024] According to the preferred embodiment the keypad 2 has a first group 7 of keys as alphanumeric keys, two soft keys 8, and a four way navigation key 10. Furthermore the keypad includes two call-handling keys 9 for initiating and terminating calls. The present functionality of the soft keys 8 is shown in a separate field in the bottom of the display 3 just above the keys 8. This key layout is characteristic of e.g. the Nokia 6210™ phone.

[0025] Figure 2 schematically shows the most important parts of a preferred embodiment of the phone, said parts

being essential to the understanding of the invention. A processor 18, which *inter alia* supports the GSM terminal software, controls the communication with the network via the transmitter/receiver circuit 19 and an antenna 20.

[0026] The microphone 6 transforms the user's speech into analogue signals; the signals formed thereby are A/D converted in an A/D converter (not shown) before the speech is encoded in an audio part 14. The encoded speech signal is transferred to the processor 18. The processor 18 also forms the interface to a RAM memory 17a and a Flash ROM memory 17b, a SIM card 16, the display 3 and the keypad 2 (as well as data, power supply, etc.). The SIM card 16 includes an electronic phonebook database 21 containing name labels and associated phone numbers. The audio part 14 speech-decodes the signal, which is transferred from the processor 18 to the earpiece 5 via a D/A converter (not shown).

[0027] A multilingual speaker independent name dialler 22 receives an audio input from the microphone 6 via the processor 18. The multilingual speaker independent name dialler 22 compares the audio input with the text strings form by the name labels in the phonebook 21 to find match and initiates that the phone calls the phone number associated with the matching name label.

Multilingual TTP Mapping

[0028] According to the invention a single TTP module 34 is used for mapping text directly into pronunciations based on a common multilingual phoneme set in the pronunciation lexicon module 32. That is, the multilingual TTP module 34 outputs a sequence of multilingual phoneme symbols based on written text as input, see Figure 4. This sequence of multilingual phoneme symbols is used for generating the pronunciation by means of a pronunciation lexicon module 32 for a multilingual recogniser 33 based on multilingual acoustic phoneme models.

[0029] The basic idea is, just as for multilingual acoustic modelling, to observe that a number of phonemes are shared among different languages. Often the same mapping between letters and phonemes exist in several languages, e.g. the letter "p" maps to the phoneme "p" (SAMPA notation) in most contexts for both English, German, Finnish and Spanish. By using the same model for similar mappings, savings in parameters can be obtained compared to using a separate TTP model for each language.

[0030] Naturally, some letters are pronounced quite differently for different languages. The ML-TTP approach (Figure 4) will therefore only be successful if the ML-TTP module 34 is capable of producing multiple, alternative phoneme symbols for such letters. The alternative phonemes are then used to create a branched grammar. The principle of branched grammar decoding in combination with a multilingual TTP module 34 is illustrated for the name "Peter" in Figure 5. The name "Peter" is pronounced quite differently in different languages, but using a multilingual TTP module 34 along with branched grammar decoding allows for capturing all pronunciations that are allowed in the set of languages supported by the multilingual recognition system. The different phonemes at each position can be weighted with probabilities as given by the multilingual TTP model, see Figure 5. This is possible when using e.g. neural networks for TTP as they can directly give estimates of the posterior probabilities of the different phonemes for each letter input.

[0031] It should be noted at this point, that when a Viterbi decoding scheme is used, only one of the possible phonemes are selected at each position during acoustic decoding. However, if an all-path forward decoder is used, all phonemes at a given position gives a weighted contribution to the score of the word, see Figure 5. Thus, a forward decoding scheme will in principle allow pronunciations that are a mixture of several different pronunciations.

Experiments

[0032] The ML-TTP method for transcribing written text into phonemes has been compared to a number of alternative TTP methods in a multilingual speaker independent name dialling task. Four languages, Finnish, German, English (US and UK), and Spanish, where used for the design and evaluation of the overall system. For these four languages, the total number of mono-phonemes is 133 corresponding to 39 phonemes for English, 28 for Spanish, 43 for German and 23 for Finnish. However, by defining a common multilingual phoneme set sharing similar phonemes, the total number of mono-phonemes can be reduced to 67 without affecting overall system performance. For testing of the overall system recognition rate a in-house test performed by the assignee of the present patent application set for each of the four languages was used. The test set for each language was based on a 120 word vocabulary of names (90 full names, 10 given names and 20 foreign names). The total number of test utterances was 21900: 5038 for UK English, 5283 for Spanish, 7979 for German and 3600 for Finnish.

[0033] A short description of the TTP, LID and multilingual acoustic model architecture, set-up, and training is provided below.

TTP mapping module

[0034] Four different approaches for TTP mapping have been considered in this work:

Table 1:

TTP training set sizes, model architectures, and model sizes. The number of input, hidden, and output units in the fully connected MLPs are denoted by I, H, and O respectively.

Language	No. words	I x H x O	Model size
UK + US	197 277	243 x 104 x 46	30 kb
German	255 280	217 x 38 x 47	10 kb
Spanish	36 486	102 x 5 x 32	0.7 kb
Finnish	15 103	72 x 4 x 25	0.4 kb
ML	882 915	333 x 99 x 73	40 kb

[0035] **TrueTTP:** true (handmade) phoneme transcriptions. This represents the "ideal" case where a lexicon covering all possible words used in the application is available.

[0036] **noLID:** language specific TTP modules assuming that the language ID of each word in the vocabulary is known a priori. The language ID can e.g. be set manually by the user or based on specific knowledge about the application.

[0037] **LID:** language specific TTP modules in combination with a statistical LID module for setting the language ID of each vocabulary word. Note that for vocabulary entries composed of several words (e.g. first and last name) the language ID is set separately for each word.

[0038] **ML-TTP:** multilingual TTP.

[0039] Instead of using a LID module 30 or assuming that the language ID is known beforehand, a pronunciation for each supported language could be generated for each word. Similarly, for some applications it makes sense to include pronunciations not only for the language selected by the LID module but also for languages known a priori to be very likely for the particular application. Such methods may, however, lead to a significant increase in real-time decoding complexity as the active vocabulary is "artificially" increased — especially when many languages are supported by the system.

[0040] There are several possible strategies for statistical TTP models including e.g. decision trees and neural networks. In this work, standard fully connected feed-forward multi-layer perceptron (MLP) neural networks have been chosen for the TTP module. The TTP networks take a symmetrical window of letters as input and gives a probability for the different phonemes for the central letter in the window. At each position in the window, the letter is encoded as an orthogonal binary vector in order to avoid introducing artificial correlations between letters.

[0041] All neural network TTP modules were designed to take up roughly the same amount of memory. Thus, the four language dependent TTP models use a total of 40 kb of memory (with 8 bit/parameter precision), which is the same amount used by the ML-TTP model. The language dependent TTP modules where trained by standard back-propagation using language specific lexicons and the ML-TTP module was trained on a balanced multilingual lexicon containing roughly equal amounts of data from each of the four languages. The size of the training databases, the architecture and the size of the TTP networks are given in Table 1. All training material was taken from the following pronunciation lexicons: BeepDic (UK), CmuDic (US), LDC-CallHome-German, LDC-CallHome-Spanish, SpeechDat-Car-Finnish transcriptions.

[0042] For all TTP methods except the TrueTTP method both single pronunciations (no branching) and branched grammars have been tested. For the branched grammars, the number of branches at each position was hard limited to a maximum of 5 and 70% of the phoneme posterior probability mass was included at each position. With this scheme, the real-time decoding complexity is increased by 10-25% corresponding to an increase of 10-25% in the number of phonemes compared to a lexicon without branching. The largest increase of 25% was observed for the ML-TTP approach. Due to pronunciation variation among different languages a larger number of branches are needed on average at each position in order to include 70%of the posterior probability mass.

LID module

[0043] As for the TTP model there are several possible choices for the statistical LID module, e.g. N-grams [1], decision trees [3], and neural networks. In a set of initial experiments, a neural network based LID module was found to have a very good generalization ability for LID classification from short segments of text even for very compact network sizes. Consequently, a standard fully connected feed-forward MLP network was selected for LID classification. The LID neural network takes a symmetrical window of letters as input and gives probabilities for each of the possible

languages at the output for the central letter in the window. The overall language probabilities for a given word was computed as a geometrical average over the language probabilities for all letters in the word. The "winning language" for a word was selected as the one with the largest overall probability.

[0044] A LID module with four outputs, 10 hidden units and 333 inputs was trained by standard back-propagation on a balanced set of 50 317 words — roughly 12 500 words from each of the four languages. All words where picked at random from the pronunciation lexicons used for training the TTP modules. With 8 bit/parameter precision, this LID module has a size of 10 kb. On an independent test set of 124 907 Finnish, German, English, and Spanish words the 10 kb LID model gives a classification rate of 86.4% on average.

Multilingual acoustic module

[0045] The acoustic phoneme models where based on a HMM/NN hybrid known as Hidden Neural Networks (HNN). The basic idea in the HNN architecture is to replace the Gaussian mixtures in each state of an HMM by state specific MLPs that have a single output and take speech feature vectors as input. A set of 67 multilingual 3-state mono-phoneme HNNs where discriminatively trained on 6965 Spanish, 9234 German, 5611 Finnish, 6300 US English, and 9880 UK English utterances taken from the Spanish-VAHA, SpeechDat-AustrianGerman, SpeechDat-Car-Finnish, Timit and WSJCAM0 databases, respectively. The number of hidden units in the network associated with a state in a particular phoneme HNN was selected based on the following heuristics: for phonemes used by a single language, zero hidden units are used. For phonemes shared by two or more languages the number of hidden units is equal to the number of languages sharing the phoneme. With 8 bit/parameter precision this results in a total size of 17 kb for the acoustic models.

[0046] Before training, the utterances where mixed with 3 different types of noise (car, café, music) at SNRs in the range 5—20dB in order to increase noise robustness of the acoustic models.

[0047] The noise mixed training files where passed through a standard MFCC preprocessor yielding 13 static, delta and delta-delta coefficient each 10 ms. Each coefficient in the 39 dimensional feature vector was normalized to zero mean and all coefficients corresponding to the log energy were normalized to unit variance.

[0048] During decoding of the HNN, a forward decoder was applied. This has been observed to yield a better performance than Viterbi decoding when the acoustic models are trained discriminatively. Furthermore, the all-path forward decoding scheme is more appropriate for branched grammars as described above.

Table 2:

Word recognition rates for various TTP methods in a multilingual speaker independent name dialling application. A single transcription is used for all entries in the vocabulary.

Single	True	noLID	LID	ML
UK English (5038)	92.6	87.8	79.7	82.2
German (7979)	95.2	92.2	86.0	87.4
Spanish (5283)	95.4	94.3	91.8	92.3
Finnish (3600)	99.1	98.9	98.5	98.3
Average	95.6	93.3	89.0	90.1

Table 3:

Word recognition rates for various TTP methods in a multilingual speaker independent name dialling application. Branched grammars are used during decoding.

Branched	True	noLID	LID	ML
UK English (5038)	92.6	91.6	81.4	85.8
German (7979)	95.2	93.5	88.7	92.5
Spanish (5283)	95.4	94.5	93.0	96.2
Finnish (3600)	99.1	98.9	98.7	98.8
Average	95.6	94.6	90.5	93.3

[0049] Even though the acoustic models have been trained using data from the above mentioned languages, they have been observed to give good performance for many other languages based on phonemes contained in the 67 multilingual phoneme set. A similar observation has been done for a HMM based multilingual system built on the same set of 67 multilingual phonemes.

Results.

[0050] The results presented below gives the overall speech recognition performance when using the TTP methods described above for transcribing text in combination with the multilingual HNN speech recogniser.

[0051] Table 2 shows the performance when using the different TTP methods for each of the four test languages when a single pronunciation is used for each entry in the vocabulary (no branched grammar decoding). The column "TrueTTP" shows the performance obtained using hand-transcriptions, "noLID", the performance obtained assuming known language ID of each entry and language specific TTP modules, and "LID" the performance obtained when using a LID module in combination with language dependent TTP modules. The last column in the table shows the performance of a system employing a multilingual TTP module.

[0052] As seen from table 2, the true transcriptions are clearly superior for all languages. However, for applications intended for portable devices, there is usually not room for storing a complete dictionary so the TTP mappings must be based on a more compact statistical method. Comparing the rows entitled "no-LID" and "LID" it is evident that the errors in language ID introduced by the statistical LID module seriously hampers the recognition performance. Thus, even with a LID module that gives a fairly accurate language identification, the performance of the overall system is seriously affected by incorrect language identification for a few words.

[0053] Table 3 illustrates the effect of applying the different statistical TTP modules with a branched grammar decoding scheme. As seen all methods gain a lot from branched grammar decoding and for the Spanish test, the ML-TTP module in combination with branched grammar decoding even outperforms the true TTP transcriptions. This indicates that an ML-TTP model allows for more variation in the pronunciation of words and thereby increases recognition performance.

Interestingly, the multilingual TTP model is capable of giving almost the same performance on average for the four languages as the combination of manually set language ID and language dependent TTP (noLID).

[0054] Table 4 illustrates the average performance over the four languages when testing in various noise environments with different TTP mapping methods. As can be seen, the gain due to branched grammar decoding and multilingual TTP is maintained in noise.

Table 4:

Average performance over Finnish, German, Spanish, and English for various TTP methods in a multilingual speaker independent name dialling application in various noise environments at 10 dB SNR. Branched grammars were applied during decoding.

Noise type	True	noLID	LID	ML
Clean	95.6	94.6	90.5	93.3
Volvo80	92.3	90.6	85.6	88.6
Babble	87.6	84.8	80.0	82.6
Pop Music	83.2	79.8	74.5	77.0

[0055] According to the invention there is provided a novel approach for generating multilingual text-to-phoneme mappings for use in multilingual speech recognition systems. The approach is based on the ability of the statistical TTP module to generate a weighting of the phonemes at each position. Here we have used the phoneme posterior probabilities provided by a feed-forward neural network, trained on data mixed from several languages.

[0056] The set of weighted phonemes at each position are used to create a branched grammar, using all-path forward decoding. The branched grammar allows for capturing of both inter- and intra-language pronunciation variations.

[0057] Tests showed that the multi-lingual TTP approach along with the branched grammar offers a significant improvement in recognition performance in multi-lingual phoneme based speaker independent speech recognition compared to using language dependent TTP modules in combination with a LID module. In some cases the multi-lingual TTP approach even improved recognition performance compared to when the language ID was known a priori.

Claims

1. A method of speech recognition comprising receiving an acoustic input and **characterised by** generating sequences of multilingual phoneme symbols based on a semantic code or code sequence, generating reference speech signals in dependence on said sequences of multilingual phoneme symbols and comparing said reference speech signals with the acoustic input in order to find a match.

2. A method according to claim 1, wherein the semantic code or code sequence represents one or more characters of an alphabet or a syllabary.

3. A method according to claim 2, wherein said code or code sequence is processed character by character and a neural network provides an estimate of the posterior probabilities of the different phonemes or phoneme sequences for each character.

4. A method according to claim 1, 2 or 3, comprising receiving a user input, generating a semantic code or code sequence in dependence on said user input, storing the generated code or code sequence in a store and retrieving said stored code or code sequence for generation of said reference speech signals.

5. A speech recognition apparatus comprising transducer means (6) for receiving an acoustic input and processing means (18, 22), **characterised in that** the processing means (18, 22) is configured for generating sequences of multilingual phoneme symbols based on a semantic code or code sequence, generating reference speech signals in dependence on said sequences of multilingual phoneme symbols and comparing said reference speech signals with the acoustic input in order to find a match.

6. An apparatus according to claim 5, wherein the semantic code or code sequence represents one or more characters of an alphabet or a syllabary.

7. An apparatus according to claim 6, wherein said code or code sequence is processed character by character and a neural network provides an estimate of the posterior probabilities of the different phonemes or phoneme sequences for each character.

8. An apparatus according to claim 7, wherein the neural network comprises a standard fully connected feed-forward multi-layer perceptron neural network.

9. An apparatus according to any one of claims 5 to 8, comprising a user input means (2), wherein the processing means (18, 22) is configured for generating a semantic code or code sequence in dependence on signals from said user input means (2) and storing the generated code or code sequence in a store (21), and for retrieving said stored code or code sequence for generation of said reference speech signals.

10. A communication terminal including an apparatus according to any one of claims 5 to 9.

11. A communication terminal according to claim 10, wherein said store (21) comprises data forming an electronic phonebook including phone numbers and associated name labels.

12. Method of speech recognition in order to identify a speech command as a match to a written text command, and comprising steps of:

providing a text input from a text database;
receiving an acoustic input;
generating sequences of multilingual phoneme symbols based on said text input by means of a multilingual text-to phoneme module;
generating pronunciations in response to said sequences of multilingual phoneme symbols; and
comparing said pronunciations with the acoustic input in order to find a match.

13. Method according to claim 12 wherein the text input is processed letter by letter, and wherein a neural network provides an estimate of the posterior probabilities of the different phonemes for each letter.

14. Method according to claim 13 comprising deriving said text input from a database containing user entered text

strings.

15. System for speech recognition and comprising:

a text database for providing a text input;
transducer means for receiving an acoustic input;
a multilingual text-to phoneme module for outputting sequences of multilingual phoneme symbols based on said text input;
pronunciation lexicon module receiving said sequences of multilingual phoneme symbols from said multilingual text-to phoneme module, and for generating pronunciations in response thereto; and
a multilingual recognizer based on multilingual acoustic phoneme models for comparing said pronunciations generated by the pronunciation lexicon module with the acoustic input in order to find a match.

16. System according to claim 15, wherein the multilingual text-to phoneme module processes said text input letter by letter, and comprises a neural network for giving an estimate of the posterior probabilities of the different phonemes for each letter.

17. System according to claim 16 wherein the neural network is a standard fully connected feed-forward multi-layer perceptron neural network.

18. System according to claim 15 wherein the text input is derived from a database containing user entered text strings.

19. System according to claim 18 wherein the database containing user entered text strings is an electronic phonebook including phone numbers and associated name labels.

20. Communication terminal having for speech recognition unit comprising:

a text database for providing a text input;
transducer means for receiving an acoustic input;
a multilingual text-to phoneme module for outputting sequences of multilingual phoneme symbols based on said text input;
pronunciation lexicon module receiving said sequences of multilingual phoneme symbols from said multilingual text-to phoneme module, and for generating pronunciations in response thereto; and
a multilingual recognizer based on multilingual acoustic phoneme models for comparing said pronunciations generated by the pronunciation lexicon module with the acoustic input in order to find a match.

21. Communication terminal according to claim 20, wherein the multilingual text-to phoneme module processes said text input letter by letter, and comprises a neural network for giving an estimate of the posterior probabilities of the different phonemes for each letter.

22. Communication terminal according to claim 21 wherein the neural network is a standard fully connected feed-forward multi-layer perceptron neural network.

23. Communication terminal according to claim 20 wherein the text input is derived from a database containing user entered text strings.

24. Communication terminal according to claim 23 wherein the database containing user entered text strings is an electronic phonebook including phone numbers and associated name labels.

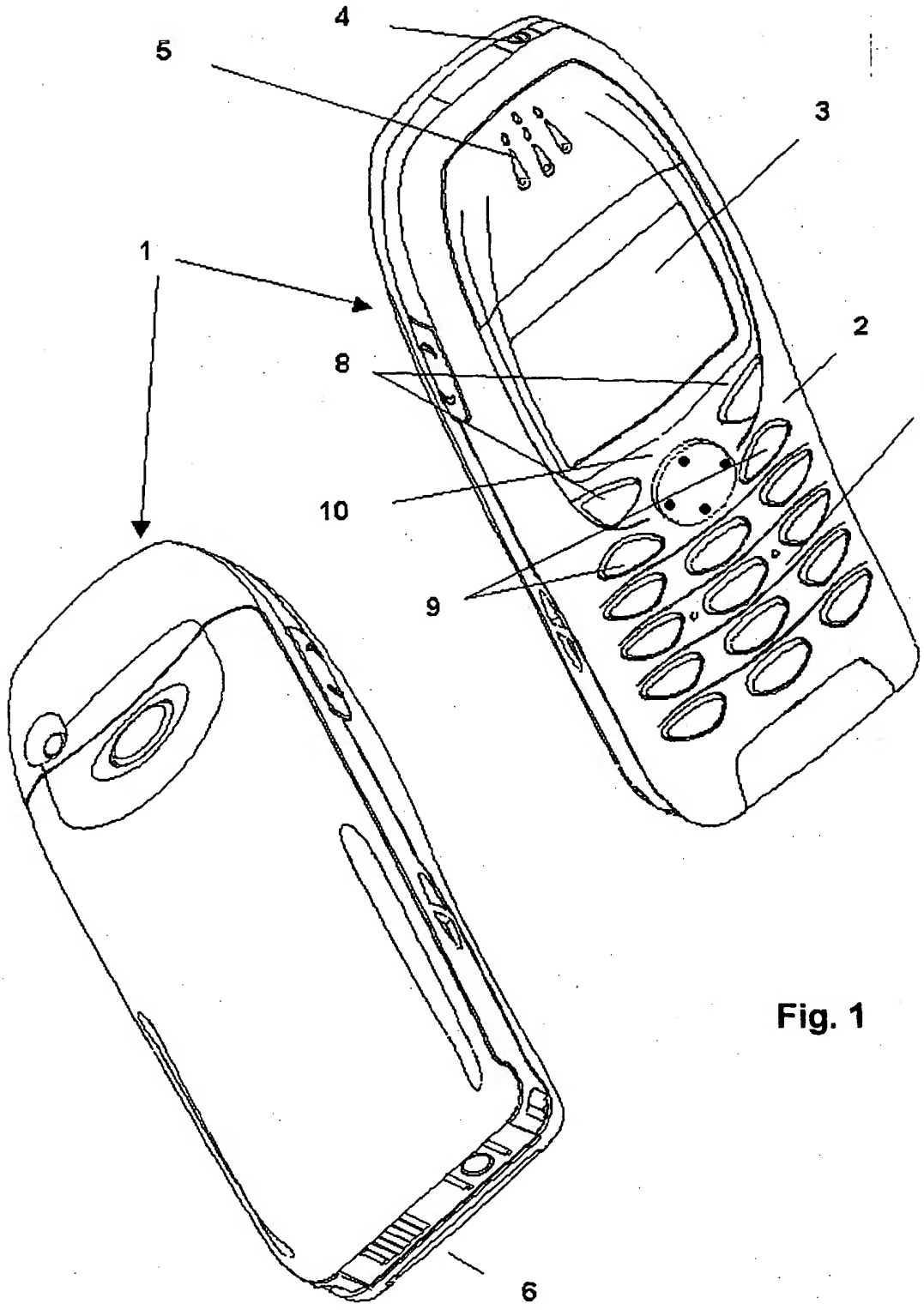

Fig. 1

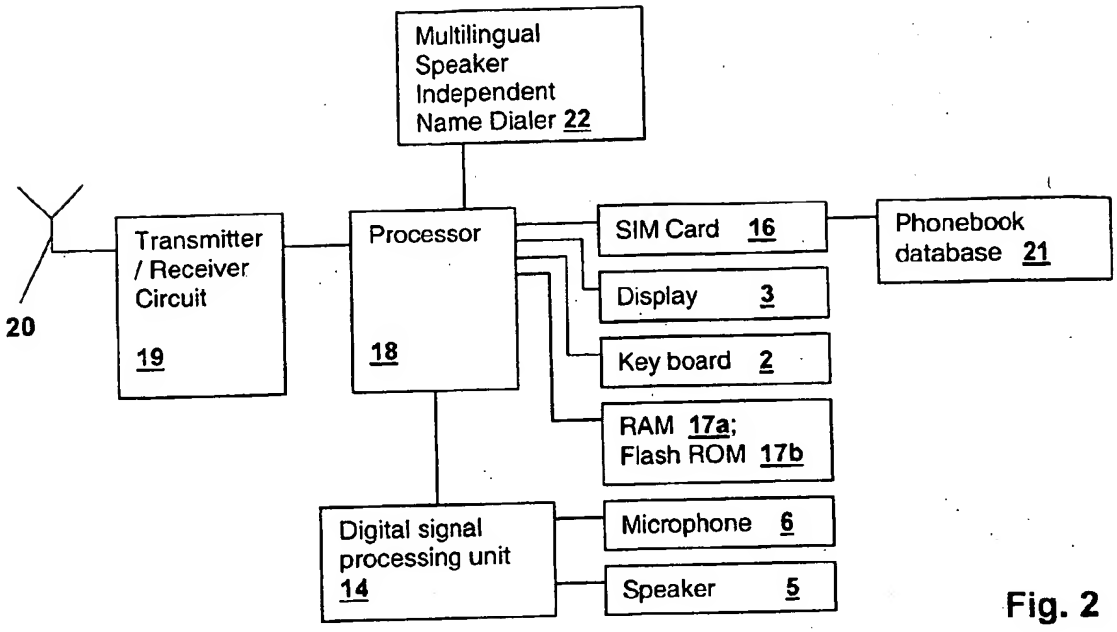

Fig. 2

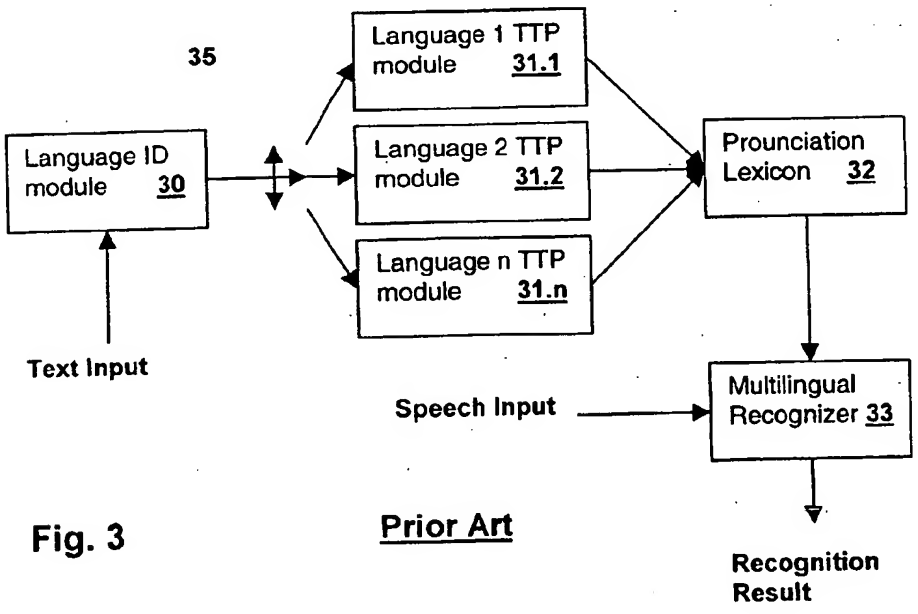

Fig. 3

Prior Art

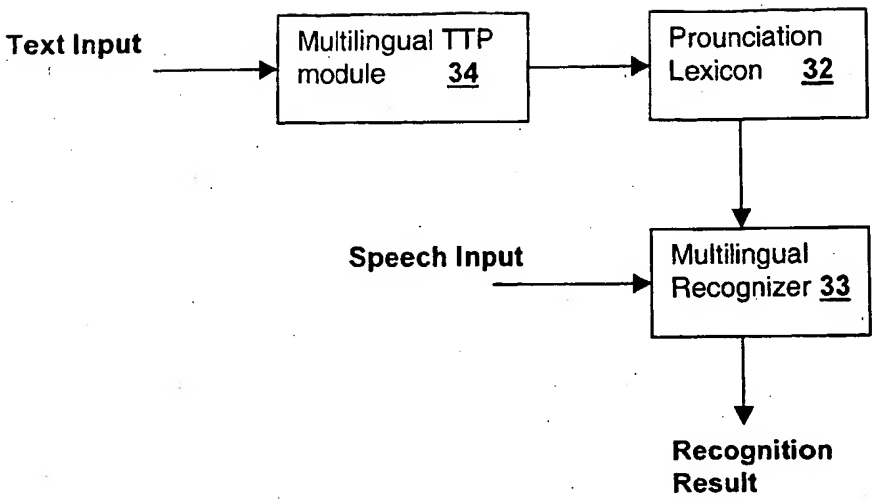
Text Input
Multilingual TTP module 34
Prounciation Lexicon 32
Speech Input
Multilingual Recognizer 33
Recognition Result

Fig. 4

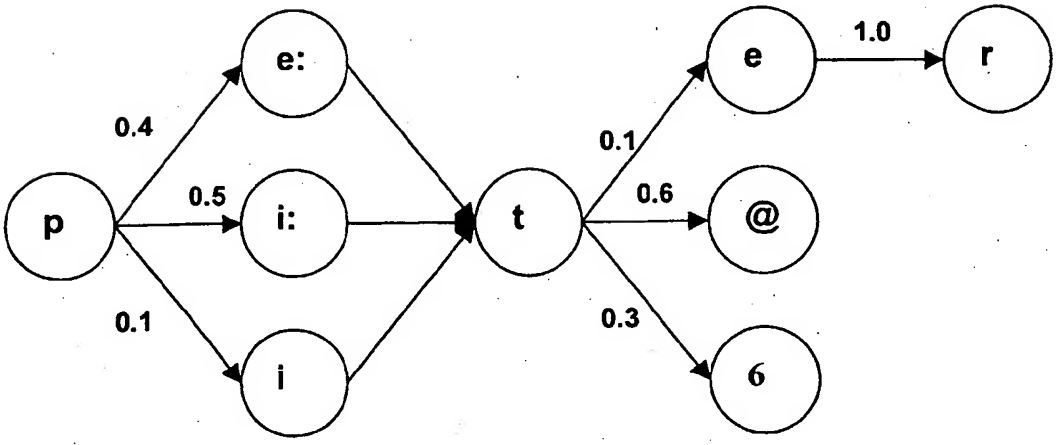
p
e:
i:
i
t
e
@
6
r
0.4
0.5
0.1
0.1
0.6
0.3
1.0

Fig. 5

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 291 848 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
02.01.2004 Bulletin 2004/01

(43) Date of publication A2:
12.03.2003 Bulletin 2003/11

(21) Application number: **02255451.3**

(22) Date of filing: **05.08.2002**

(51) Int Cl.[7]: **G10L 13/08**, G10L 15/14

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **31.08.2001 US 942609**

(71) Applicant: **Nokia Corporation**
02150 Espoo (FI)

(72) Inventors:
- **Riis, Soeren**
  2750 Ballerup (DK)
- **Jensen, Kare Jean**
  2830 Virum (DK)
- **Pedersen, Morten With**
  2100 Koebenhavn (DK)

(74) Representative: **Geary, Stuart Lloyd et al**
Venner, Shipley & Co.,
20 Little Britain
London EC1A 7DH (GB)

(54) Multilingual pronunciations for speech recognition

(57) There is provided a novel approach for generating multilingual text-to-phoneme mappings for use in multilingual speech recognition systems. The multilingual mappings are based on the weighted outputs from a neural network text-to-phoneme model, trained on data mixed from several languages. The multilingual mappings used together with a branched grammar decoding scheme is able to capture both inter- and intra-language pronunciation variations which is ideal for multilingual speaker independent speech recognition systems. A significant improvement in overall system performance is obtained for a multilingual speaker independent name dialling task when applying multilingual instead of language dependent text-to-phoneme mapping.

EP 1 291 848 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 02 25 5451

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	JILEI TIAN ET AL: "Pronunciation and Acoustic Model Adaptation for Improving Multilingual Speech Recognition" INTERNATIONAL WORKSHOP ON ADAPTATION METHODS FOR SPEECH RECOGNITION, 29 - 30 August 2001, pages 131-134, XP007005515 Sophia Antipolis, France	1,2,4-6, 9-12,15, 20	G10L13/08 G10L15/14
Y	* the whole document *	3,7,8, 13,14, 16-19, 21-24	

Y	DESHMUKH N ET AL: "An advanced system to generate pronunciations of proper nouns" ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1997. ICASSP-97., 1997 IEEE INTERNATIONAL CONFERENCE ON MUNICH, GERMANY 21-24 APRIL 1997, LOS ALAMITOS, CA, USA,IEEE COMPUT. SOC, US, 21 April 1997 (1997-04-21), pages 1467-1470, XP010226082 ISBN: 0-8186-7919-0 * abstract * *section 4.4 Multilayered Perceptron Network *	3,7,8, 13,14, 16-19, 21-24	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G10L

P,X	RIIS S.K., PEDERSEN M.W. AND JENSEN K.J.: "Multilingual Text-To-Phoneme Mapping" EUROSPEECH 2001, vol. 2, 3 - 7 September 2001, page 1441 XP007004620 Aalborg, Denmark * the whole document *	1-24	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	3 November 2003	Quêlavoine, R

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☑ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

THIS PAGE BLANK (USPTO)